Fast Spatio-Temporal Data Mining from Large Geophysical Datasets

P. Stolorz*, E. Mesrobian+, R. Muntz+, J.R. Santos+,
E. Shek+, J. Yi+, C. Mechoso** and J. Farrara**

*Jet Propulsion Laboratory, California 1 nstitute of Technology

+ Computer Science Department, (University of California, Los Angeles

** Department of Atmospheric Sciences, University of California, Los Angeles

Abstract

The scientific challenge of understanding global climate change requires the application of knowledge discovery and datamining techniques on a large scale. Advances in parallel supercomputing technology enable high- resolution modeling, while sensor technology allows data capture on an unprecedented scale. We discuss here experiences with a data analysis environment developed at UCLA, CONQUEST, which provides content-based access to such scientific datasets. CONQUEST (CONtent-based Querying in Space and Time) employs a combination of workstations and massively parallel processors (MPP's) to mine geophysical datasets possessing a prominent temporal component. It is designed to enable complex multi-] odal interactive querying and knowledge discovery, while simultaneously coping with the extraordinary computational demands posed by the scope of the datasets involved. We review its use to perform automatic cyclone extraction and detection of spatio-temporal blocking conditions on MPP platforms.

Submitted to:
1st International Conference on Knowledge Discovery and Data Mining (1<1111-95)
August 20-21, 1995

1 Introduction

Understanding the long-term behavior of the earth's atmospheres and oceans is one of a number of ambitious scientific and technological challenges which have been classified as " Grand Challenge" problems. These problems share in common the need for the application of enormous computational resources if they ale to be solved. Substantial progress has of course already been made on global climate analysis over the years, due on the one hand to the development of ever more sophisticated sensors and data-collection devices, and on the other to the implementation and analysis of large-scale models on supercomputers. Gigabytes of data can now be generated with relative case for a variety of important geophysical variables over long time scales. However, this very success has created a new problem: how do wc store, manage, access and interpret the vast quantities of information now at our disposal?

The issue of data management and analysis is in itself a Grand Challenge which must be addressed if the production of real and synthetic data on a large scale is to prove truly useful. The challenge has been addressed by the development at UCLA of of CONQUEST (CONtent-based QUerying in Space and Time) [1], a distributed parallel querying and analysis environment developed to address this challenge in a geoscientific setting. The basic idea of CONQUEST is to supply a knowledge discovery environment which allows geophysical scientists to 1) easily formulate queries of interest, especially the generation of contmt-based indices dependant on both "specified" and "emergent" spatio-temporal patterns, 2) execute these queries rapidly on massive datasets, 3) visualize the results, and 4) rapidly and interactively infer and explore new hypotheses by supporting complex compound queries (in general, these queries depend not only on the different datasets themselves, but also on content-based indices supplied by the answers to previous queries).

Contmt-based access to image databases is a rapidly developing field with applications to a number of different scientific, engineering and financial problems. A sampling may be found in volumes such as [2, 3]. One example is the QUBIC project [4] illustrating the state-of-the-art in image retrieval by content, while examples of work in the area of geoscience databases include JARTool [5], VIMSYS [G] and Sequoia 2000 ['/]. Many of these efforts are directed at datasets which contain relatively static l]igh-resolution spatial patterns, such as high-resolution Landsat imagery, and Synthetic Aperture Radar imagery of the earth's surface and of other planets. CONQUEST shares a great deal in common with these systems. Its distinguishing features arc, 1) the fact that it is designed to address datasets with prominent temporal components in addition to significant high-resolution spatial information, and 2) that it is designed from the beginning to take maximum advantage of parallel and distributed processing power.

2 System Architecture

The system architecture is outlined in Figure 1. Details can be found in []]. It consists of the the following 5 basic components:

- Scientist Workbench
- Query Manager (parser and optimizer)
- Visualization]) Manager
- Query execution engine
- In formation Repository

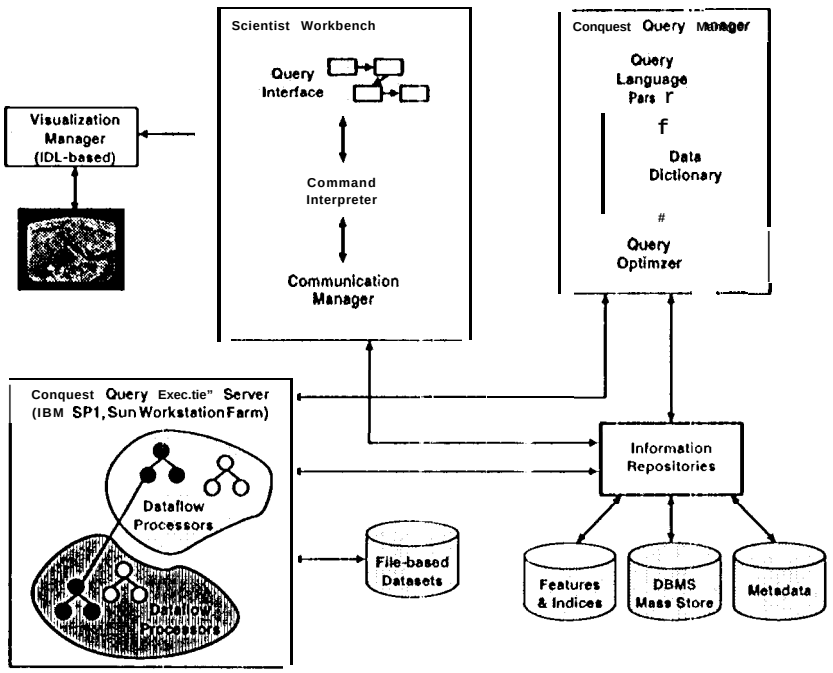

Figure 1: Sytem Archit ecture

The scientific workbench consists of a graphical user interface enabling the formulation of queries in terms of imagery presented on the screen by the Visualization Manager. Queries formulated on the workbench arc parsed aud optimized for target architectures by the Query Manager, and then passed onto the execution engines. These can be either parallel or serial super computers, such as IBM SP1 and Intel Paragon supercomp uters, single workstations, or workstation farms. The simplest queries consist of the extraction of well-dcflncd features from "raw" data, without reference to any other information. These features arc registered with the information Depository to act as indices for further queries. Salient information extracted by queries can also be displayed via the Visualization Manager. The latter is implemented on top of IDL, and supports static plotting (21) and 3D graphs) of data, analysis of data (e. g., statistical, contours), and animation of datasets.

3 Datasets

CONQUEST has been applied to datasets obtained from two different sources. The first dataset is output from an Atmospheric Global Circulation Model developed at UCLA, chosen for two principal reasons: (1) it includes a challenging set of spatial-temporal patterns (e.g., cyclones, hurricanes, fronts, and blocking events); and (2) it is generally free of incomplete, noisy, or contradictary information. The UCLA atmospheric general circulation model (AGCM) [8] is a finite-difference model that includes sophisticated parameterizations of cumulus convection [9, 1 O], as well as planetary boundary layer processes and parameterizations of shortwave and longwave radiative transfer. The horizontal structure of the model is typically represented by grid cells of various resolution; we are using a grid size of 5° longitude and 4° of latitude. The vertical component of the model is represented by a series of pressure layers; the version in this study has 9 layers in the vertical with the top at 50 millibars.

The prognostic variables of the AGCM are horizontal velocities, potential temperature, water vapor and ozone mixing ratio, surface pressure, ground temperature, and the depth of the planetary boundary layer. There are also diagnostic variables such as vertical velocities, precipitation, cloudiness, surface fluxes of sensible and latent heat, surface wind stress and radiative heating. Typically, the model's output is written out to the database at 12-hour (simulation time) intervals; however, this frequency can be modified depending on storage capacity of the database. the model can be run with different spatial resolutions (grid sizes) and temporal resolution (output frequency). At the lowest spatial resolution (4° x 5°, 9 levels) with 12 hour output interval, the AGCM produces approximately 5 Gbytes of data per simulated year, while a 100-year simulation of a AGCM with a 1° x 1.25°, 57 levels) generates approximately 30 terabytes of output.

The second dataset is obtained from ECMWF (European Center for Medium-range Weather Forecasting), and is split into two subgroups based upon simulated data and satellite data respectively. The ECMWF T42L19 and T42L19 VIIp AMIP 10 Year Simulation (1979-1988) dataset contains fields with a grid size 128 longitudinal points (2.81 25°) by 64, Gaussian latitudinal points, by 15 pressure levels. Model variables were output to files every 6-hour (simulation time) intervals. Each 4D variable (e.g., geopotential height) requires 7Gb of disk storage. The ECMWF TOGA Global Basic Surface and Upper Air Analyses dataset consists of fields which are uninitialized analyses sampled twice a day (O GMT and 1200 GMT), at 14 or 15 pressure levels, over a (2.5° longitude by 2.5° latitude grid. Upper air variables include geopotential, temperature, vertical velocity, u- and v- components of horizontal wind, and relative humidity, while surface variables include surface pressure, surface temperature, mean sea-level pressure, etc.. The dataset requires about 130 Mb/month.

4 Spatio-temporal. Feature Extraction

We review here the use of CONQUEST to capture heuristic rules for prominent features, as discussed in [1]. Two canonical features are cyclones and blocking features. These phenom-

ena interact in a manner that is still imperfectly understood, and therefore represent ideal candidates for the implementation of complex queries.

4.1. Cyclone detection

Cyclones are some of the most prominent climatic features displayed by Global Circulation Models. There is, however, no single objective definition in the literature of the notion of a cyclone. Several working definitions are based upon the detection of a threshold level of vorticity in quantities such as the atmospheric pressure at sea level. others are based upon the determination of local minima of the sea level pressure [11]. The latters' careful treatment includes the introduction of extra relevant information such as prevailing wind velocities in a meaningful way.

Cyclones are defined below as one-dimensional tracks in a 3-dimensional space consisting of a time axis and the 2 spatial axes of latitude and longitude. Cyclones represent paths of abnormally low sea level pressure in time. A typical cyclone track, in this case over the continental United States, is shown schematically in Figure 2, together with a dataflow description of the associated cyclone query. The track is found by first detecting onc or more local minima in the 2-dimensional grid of sea level pressure values representing a single time, frame of the GCM. A local minimum is found by locating a grid location whose pressure value is lower than that at all the grid points in a neighborhood around the location by some (adjustable) prescribed threshold. This minumum is then refined by interpolation using low-order polynomials such as hi-cubic splints or quadratic bowls. Given a local minimum occurring in a certain GCM frame, the central idea is to locate a cyclone track by detecting in the subsequent GCM frame a new local minimum which is "sufficiently close" to the current one. Two minima are deemed "sufficiently c1osc" to be part of the same cyclone track if they occur within 1/2 a grid spacing of each other. Failing this condition, they are also "sufficiently c1osc" if their relative positions are consistent with the instantaneous wind velocity in the region. A trail of several such points computed from a series of successive frames constitutes a cyclone.

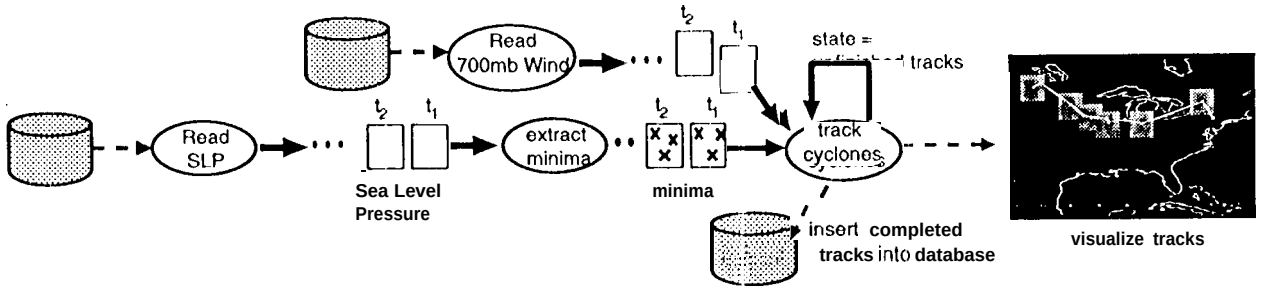

Figure 2: 1)ata flow representation of the cyclone tracking query.

Figures 3 and 4 present cyclopresense density maps of cyclones during the northern winter extracted from-ECMWF model and analyses datasets, respectively. In the figures, white represents the lowest density value, while black indicate- the largest density value.

In the ECMWF analyses ("observational") dataset (Figure 3), as in the real atmosphere, most of extratropical cyclones arc formed and migrate within a few zonally-elongated regions (i.e., "stormtracks") in the northern Atlantic and Pacific and off around the Antarctic. The ECMWF AGCM (4), however, tends to yield significantly more cyclones than observed.

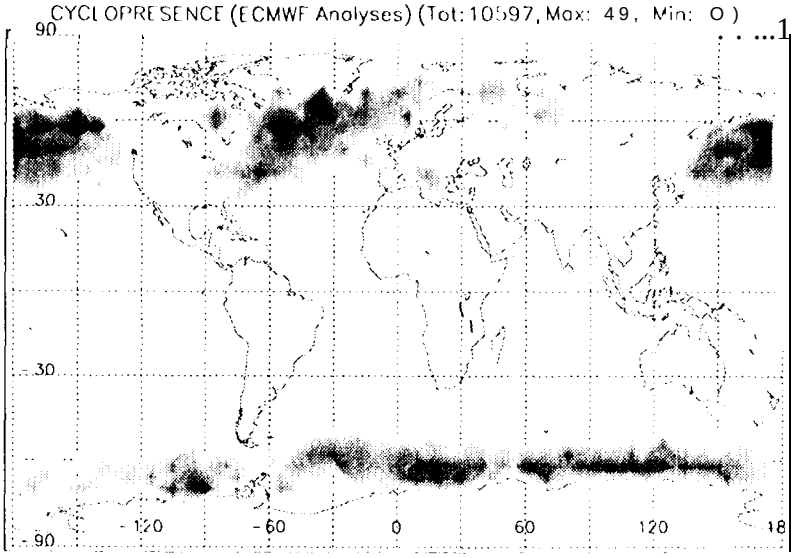

Figure 3: Cyclopresense density map of cyclones during the nothern winter extracted from the ECMWF Analyses dataset (1 985-1994).

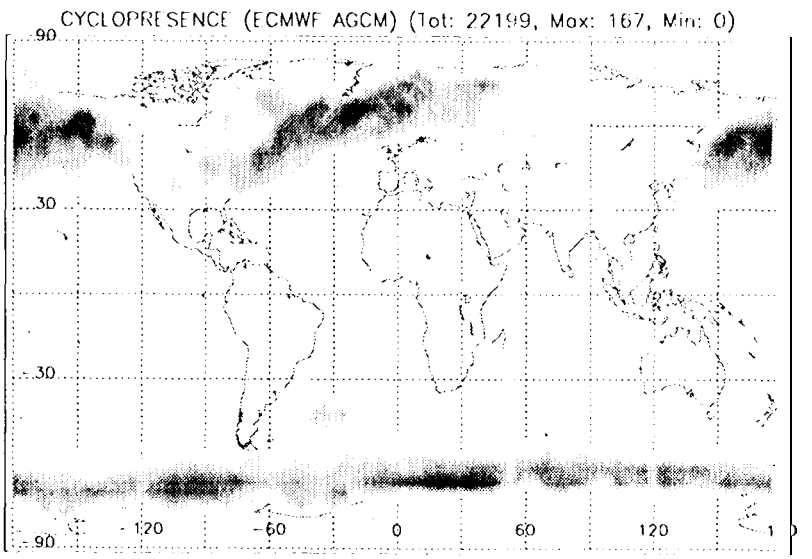

Figure 4: Cyclopresense density map of cyclones during the nothern winter extracted from the ECMWF GCM model data (1 979-1988).

4.2 Blocking Feature extraction

On time scales of one to two weeks the atmosphere occasionally manifests features which have well-defined structures and exist for an extended period of time essentially unchanged in form. Such structures arc referred to, in general, as "persistent anomalies". Onc particular class of persistent anomalies, in which the basic westerly jet stream in mid-latitudes is split into two branches, has traditionally been referred to as "blocking" events. The typical anomalies in surface weather (i.e., temperature and precipitation) associated with blocking events and their observed frequency have made predicting their onset and decay a high priority for medium-ran,gc (5-15 day) weather forecasters.

While there is no general agreement on how to objectively define blocking events, most definitions require that the following conditions exist: 1) the basic westerly wind flow is split into two branches, 2) a large positive geopotential height anomaly is present downstream of the split, and 3) the pattern persists with recognizable continuity for at least 5 days. Blocking features arc determined by measuring the difference between the geopotential height at a given time of year and the climatological mean at that time of year am-aged over the entire time range of the dataset. Before taking this difference, the geopotential height is first passed through a low-pass temporal filter (a 4th order Butter worth filter with a 6-day cut-off), to ensure that blocking signatures are not contaminated by the signals of migratory cyclones and anticyclones. The filtered field is averaged to obtain the mean year. A Fourier transform of the mean year is then taken, followed by an inverse Fourier transform on the first four Fourier components. This procedure yields smooth time series for seasonal cycles even if the dataset is small ($<\approx$ 100 years). The resulting filtered mean year is subsequently compared with the Iluttcrworth-processed geopotential height fields to generate the fundamental anomaly fields. Blocking "events" can be detected as time periods δt during which filtered geopotential anomaly values are persistently higher than θ. Figure 5 presents a density plot indicating the global occurrances of Mocking events for UCLA AGCM data (1985-1989), extracted using δt = 5 days and $\theta = 0.5\sigma$. In the figur c, white represents -he lowest density value, while black indicates the largest density value. Since 1)locking is by nature an extratropical phenomenon, wc have eliminated values in the tropics from the plot.

4.3 Parallel Implementation of Feature Detection

The algorithms described above for extracting cyclone and blocking features on a 10-year dataset of atmospheric data require several hours to execute on a typical scientific workstation. Pre-processing and storage of indices by workstations is of course a feasible alternative for heavily used features, but will not suffice for a more general and wide-ranging querying capability. 1 t is here that massively parallel processors (MPP's) enter the picture. The features described above can be computed quite efficiently on MPP's, bringing the turn-around time for a typical query down to the range of minutes on mcdiul]l-scale parallel machines that have been used to date (a 24-node IBM S1'1 and a 56--node Intel Paragon). It is expected that near real-tilllc performance will be achieved when the system is ported to larger platforms comprising up to 512 nodes.

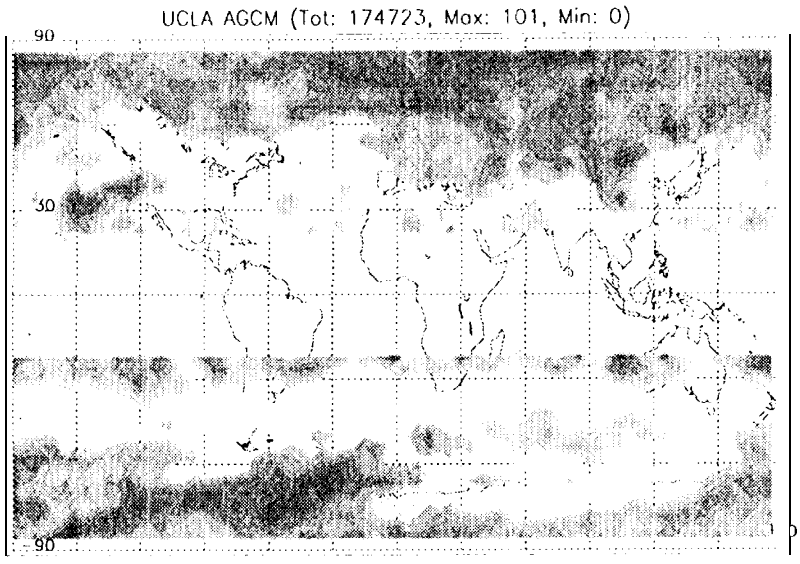

Figure 5: Density map of blocking events extracted from the UCLA AGCM model data-(1985-1989).

The parallel implementation of these queries requires of course an explicit decomposition of the problem across the various nodes of a parallel machine. In the case of cyclone detection, the optimal decomposition is based upon a division of the problem into separate temporal slices, each of which is assigned to a separate node of the machine. A temporal decomposition such as this proves to be highly efficient on a coarse-graincd architecture, provided that cyclone results obtained during a given time zone do not interfere too strongly with those at a later time. Care must be exercised in such a decomposition, as the temporal dimension does not typically parallelize in a natural way, especially when state information plays an important role in the global result. State-informatioll plays a fundamental role in the very definition of cyclones, so care must obviously be taken in the ensuing parallel decomposition. The problem proves tractable in the case of cyclone detection because of the observation that no cyclones last longer than 24 frames. This allows the use of a straightforward temporal shadowing procedure, in which each node is assigned a small number of extra temporal frames that overlap with the first fcw frames assigned to its successor node. In the case of blocking feature detection, a straightforward spatial decomposition which assigned different Mocks of grid points to different machine nodes proves to be optimal.

5 Conclusions

Extensible query processing systems in which scientists can easily construct content-based queries have been reviewed that enable important features present in geophysical datasets to be extracted aud catalogued efficiently. Examples include cyclone tracks and Mocking events from both observational and simulated datasets on the or dcr of gigabytes in size.

Several future issues must be addressed by reseal chers in the field. Onc is the popu-

lation of the query set with a wider range of phenomena including oceanographic as well as atmospheric queries. Another is the application of machine learning methods to extract previously unsuspected patterns of interest. A third issue is the scaling of system size onto massively parallel platforms, a necessary ingredient to cope with the terabyte size datasets that are becoming available. in this area, scaleable I/O considerations are at least as important as those associated with computation *per* se, and are an active area of research. A final issue is the development of an appropriate field- model language capable of expressing queries based upon large imagery datasets rapidly and efficiently.

Acknowledgements

The research described in this paper was carried out by the Jet l'repulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. Wc thank J. Sphar for supplying the AGCM output and C.C. Ma for valuable discussions on cyclogenesis.

References

[1] E.C. Shck and R.R. Muntz. The Conquest Modeling Framework for Geoscientific Data. In *UCLA Technical Report*, 1994.

[2] In *Visual Database Systems*. North Holland, 1992.

[3] S-K. Chang and A. Hsu. Image information systems: Where do wc go from here? *IEEE Transactions on Knowledge* and *Data Engineering*, 4(5):431- 442, 1992.

[4] W. Niblack, R. Barber, and W. ct. al. Equitz. The qubic project: Querying images by content using color, texture, and shape. Technical Report Research Report #RJ 9203, 1 BM Research Division, Feb. 1993.

[5] U. M. Fayyad, P. Smyth, N. Weir, and S. Djorgovski. Automated analysis and exploration of large image databases: results, progress, and challenges. *Journal of Intelligent Information Systems*, 4:1-19, 1994.

[6] A. Gupta, T. Weymouth, and R. Jain. Semantic queries with pictures: The vymsys model. in *Proceedings of VLDB, Barcelona, Spain*, pages 69-'79, Scpt. 1991.

[7] A. Guptill and M. Stonebraker. The sequoia 2000 appi each to managing large spatial object databases. In *Proc. 5th Int'l. Symposium on Spatial Data Handling, Charleston, S. C.*, pages 642-651, Aug. 1992.

[8] C. R. Mechoso, S. W. Lyons, and J. A. Spahr. The impact of sca surface temperature anomalies on the rainfall over northeast brazil. *J. Clim.*, 3:812- 826, 1990.

[9] A. Arakawa and W. H. Schubert. Interaction of a cumulus cloud ensemble with the large-scale environment, part i. *J. Atmos. Sci.*, 31:674- 701, 1974.

[10] S. J. Lord, W. C. Chao, and A. Arakawa. Interaction of a cumulus cloud ensemble with the large-scale environment. part iv: The discrete model. *J. A tmos. Sci.*, 39: 104–1 13, 1982.

[11] R.J. Murray and I. Simmonds. A numerical scheme for tracking cyclone centres from digital data. part i: development and operation of the scheme. *Aust. Met. Mag*, *39:155*-166, 1991.